Unsupervised Morphology-Based Vocabulary Expansion

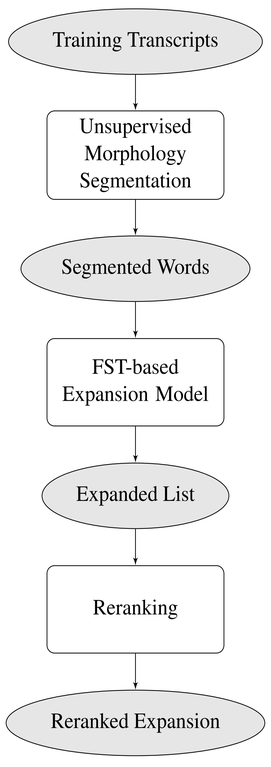
Mohammad Sadegh Rasooli, Thomas Lippincott, Nizar Habash, and Owen Rambow

Center for Computational Learning Systems, Columbia University

{rasooli,tom,habash,rambow}@ccls.columbia.edu

Introduction

- **Objective**
  - Creating new words to extend vocabularies for under-resourced languages
- **Approach**
  - Using unsupervised learning of morphology and using learned affixes to generate new words
- **Tools**
  - Morfessor for unsupervised Segmentation (Creutz and Lagus, 2007)
  - WFSTs for generating new words



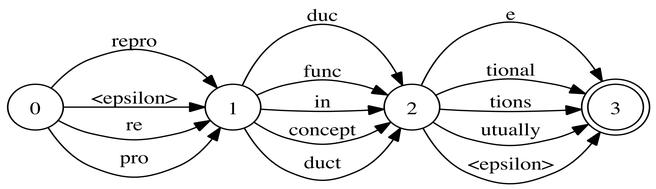
Flowchart of the vocabulary expansion model

Modeling Word Generation

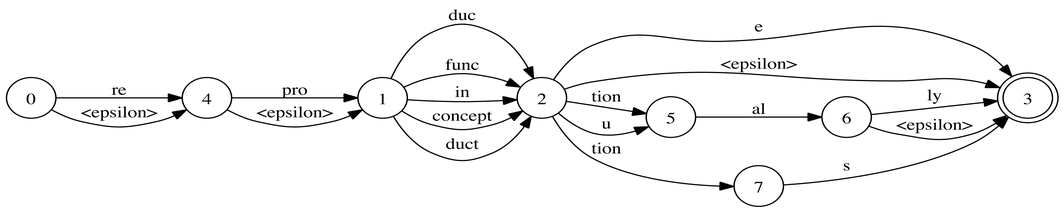
- **Two Word Models for Expansion**
  - **Fixed Affix**: every word is a sequence of one/zero complex (multi-morpheme) prefix, stem and one/zero complex (multi-morpheme) suffix.
  - **Bigram Affix**: every word only has one stem and zero or more morpheme affixes.
- **Reranking Models**
  - Reranking with letter trigraph probabilities
  - Reranking with letter trigraphs at morpheme boundaries only
  - No Reranking

```
re+ pro+ duc +e
func +tion +al
re+ duc +e
re+ duc +tion +s
in
pro+ duct
concept +u +al + ly
```

(a) Training data with morpheme boundaries. Prefixes end with and suffixes start with "+" signs.



(b) FST for the Fixed Affix expansion model

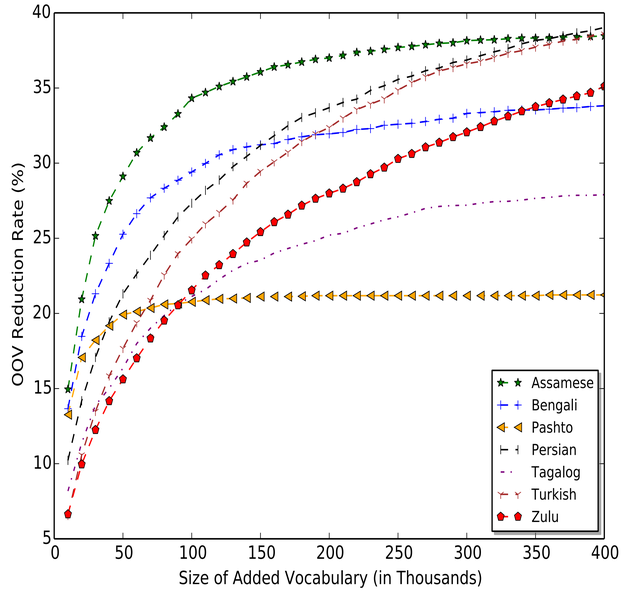


(c) FST for the Bigram Affix expansion model

Two models of word generation from morphologically annotated data

Experiments and Results

- We ran Morfessor on 65K to 115K tokens from seven different languages
- We evaluated on a small-sized data set (50K to 100K tokens) measuring out-of-vocabulary reduction.
- The best results use the Fixed Affix model with trigram re-ranking.
- Word precision is still a big issue (less than 30% of the top 50K generated types could be analyzed by a Turkish morphological analyzer).



Token-based OOV reduction with different expansion sizes for the Fixed Affix model with trigraph reranking